ELSEVIER

Artificial Intelligence 107 (1999) 347–356

Artificial Intelligence

Research Note

Adaptive regularization parameter selection method for enhancing generalization capability of neural networks

Chi-Tat Leung [1], Tommy W.S. Chow *

Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

Received 13 July 1998; received in revised form 27 October 1998

Abstract

A novel adaptive regularization parameter selection (ARPS) method is proposed in this paper to enhance the performance of the regularization method. The proposed ARPS method enables a gradient descent type training to tunnel through some of the undesired sub-optimal solutions on the composite error surface by means of changing the value of the regularization parameter. Undesired sub-optimal solutions are introduced inherently from regularized objective functions. Hence, the proposed ARPS method is capable of enhancing the regularization method without getting stuck at these sub-optimal solutions. © 1999 Published by Elsevier Science B.V. All rights reserved.

Keywords: Neural network; Regularization method; Generalization capability

1. Introduction

Neural networks are a burgeoning area of artificial intelligence and are applied in many engineering applications, such as time-series forecasting, and signal processing. The mean squares (MS) error function is used extensively in the training of backpropagation neural networks (BPnet). Until now, most of the fast learning algorithms were derived based on the MS error function. Despite the popularity of the MS error function, there are two main shortcomings in applying those MS error based algorithms for general applications. On the one hand, there are many sub-optimal solutions on the MS error surface. The network training may easily stall because of being stuck in one of the sub-optimal solutions. On

* Corresponding author. Email: eetchow@cityu.edu.hk.

[1] Current address: Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. Email: chitat@en.polyu.edu.hk.

0004-3702/99/$ – see front matter
PII: S0004-3702(98)00115-5

the other hand, the MS error function, in general, is a universal objective function to cater all harsh criteria of different applications. However, there is a common view that different applications may emphasize on different aspects. To have an optimal performance such as a low training error and high generalization capability, additional assumptions and heuristic information on a particular application have to been included. As BPnets are considered as universal approximators [3,5], the network training is a process of non-parametric functional estimation in the statistical sense. The network training based on a finite number of training examples is basically an ill-posed problem [6]. However, there is only a few number of the functional variances to have high generalization capability. Our main concern is that the error level for the novice examples should be comparable to that of the training data in the particular application. In other words, the trained BPnets should have a high degree of generalization capability and in order that the training error can be considered as a reliable measure of the network performance. In order to have such an optimal network, the inclusion of *a priori* knowledge can facilitate the network training to converge to a desirable functional estimate [1]. One of the techniques to absorb the *a priori* knowledge is regularization. Although the regularization technique is a systematic approach to make the network training less ill-posed, the training process may stall due to the existence of sub-optimal solutions with the newly constructed objective function. This paper addresses this undesirable stalling problem in the regularization method and proposes an adaptive regularization parameter selection (ARPS) method to enhancing the network performance.

2. Regularization method

Recently, a number of techniques have been proposed to include heuristic information and assumptions in the training process for different applications. One of the approaches can be coarsely classified as the regularization technique. The technique constructs a regularized objective function to assimilate the *a priori* knowledge. For example, in the applications of classification, the new discriminant functions [9] were proposed to maximize the classification accuracy for the unseen examples. For the applications of functional approximation, a number of regularized objective functions have been proposed to enhance the generalization capability. For instance, the regularized objective functions are derived in accordance with the techniques such as searching for flat minima [7], minimizing the mutual information criterion [4], and minimizing the higher-order cumulants between the network output and the desired output [8]. Although the regularization method is a systematic approach to make the network training less ill-posed, the training process may stall due to the existence of sub-optimal solutions of the newly constructed objective function.

We consider a typical form of the regularized objective function which is expressed in the following equation:

$$H(\boldsymbol{W}, \boldsymbol{D}) = M(\boldsymbol{W}, \boldsymbol{D}) + \lambda P(\boldsymbol{W}, \boldsymbol{D}), \tag{2.1}$$

where $\boldsymbol{W} = (w_0, w_1, \ldots, w_m)^{\mathrm{T}}$ is the weight vector of the BPnet; $\boldsymbol{D}$ is the set of training examples; λ is the regularization parameter and the superscript T denotes the matrix

transpose operation. $M(\boldsymbol{W}, \boldsymbol{D})$, which is mostly the MS error, is the primary cost term; and $P(\boldsymbol{W}, \boldsymbol{D})$ is the regularization term which is used to assimilate the *a priori* knowledge. For example, when weight decay method [2] is used, the regularization term $P(\boldsymbol{W}, \boldsymbol{D})$ will be $\sum_i w_i^2$. In general, the plausible range of λ is determined experimentally. The value of λ is often pre-selected within its plausible range or λ is selected in accordance with some heuristic selection schemes. It is believed that a systematic λ selection mechanism may be able to further enhance the generalization capability of the trained BPnets. In fact, this type of the objective function is still suffering from the problem of existing sub-optimal solutions due to the nonlinearity of BPnets. Although some undesirable solutions should be screened out to some extent, the regularized objective function introduces another set of undesirable solutions. Consequently, the enhancement in the regularization technique may sometimes be insignificant, especially when a fixed value of λ is used during the network training.

From Eq. (2.1), the sub-optimal and optimal solutions are reached only when the gradient of $H(\boldsymbol{W}, \boldsymbol{D})$ is a zero vector, viz. $\nabla H(\boldsymbol{W}, \boldsymbol{D}) = \boldsymbol{0}$.

(1) The $\nabla M(\boldsymbol{W}, \boldsymbol{D})$ and $\nabla P(\boldsymbol{W}, \boldsymbol{D})$ are both zero vectors;

(2) The $\nabla M(\boldsymbol{W}, \boldsymbol{D})$ and $\nabla P(\boldsymbol{W}, \boldsymbol{D})$ are both nonzero vectors such that

$$\nabla M(\boldsymbol{W}, \boldsymbol{D}) + \lambda \nabla P(\boldsymbol{W}, \boldsymbol{D}) = \boldsymbol{0}, \tag{2.2}$$

where ∇ is the gradient operator with respect to $\boldsymbol{W}$. Condition (1) is a trivial case. The standard network training is often expected to converge to the minimum of this condition. Condition (2) may contribute to the introduction of another set of undesirable sub-optimal solutions. Also, the location of the sub-optimal solutions of condition (2) is significantly affected by the pre-selected value of λ. Hence, the selection of λ is one of the major issues in the regularization techniques and is determinant in the performance of BPnets, especially for the generalization capability.

3. Adaptive regularization parameter selection method

As the selection of λ is extremely crucial to the performance of BPnets, this paper proposes a novel adaptive λ selection mechanism. Our ARPS method has been divided into three main elements according to its functionality. The three functional elements are responsible for the following functions:

- *Stalling identification method* identifies whether the training process converges to a sub-optimal solution that satisfies condition (2).
- λ *selection scheme A* selects an appropriate value of λ to ensure the training convergence of the $M(\boldsymbol{W}, \boldsymbol{D})$ and $P(\boldsymbol{W}, \boldsymbol{D})$ when the training process is not stuck into a sub-optimal solution that satisfies condition (2).
- λ *selection scheme B* selects an appropriate value of λ to ensure the training convergence of the $M(\boldsymbol{W}, \boldsymbol{D})$ when the training process may stall in the sub-optimal solution.

On the one hand, based on their functions, the λ selection scheme A guarantees the convergence of the terms $M(\boldsymbol{W}, \boldsymbol{D})$ and $P(\boldsymbol{W}, \boldsymbol{D})$ when there is no clue indicating that the training process will stall. This part can assure that the training process goes as smooth

as possible. On the other hand, the λ selection scheme B will be applied to prevent the network training from stalling at a sub-optimal solution satisfying condition (2) when the ARPS method identifies the training process to be about to stall at the sub-optimal solution. Hence, within the plausible range of the λ, the ARPS method is capable of avoiding the training process from stalling at a sub-optimal solution of condition (2). The detailed descriptions of the stalling identification method and the two λ selection schemes are given in the sections below.

3.1. Stalling identification method

According to Eq. (2.2), the stalling situation of condition (2) occurs when the vector sum of the nonzero terms $\nabla M(\boldsymbol{W}, \boldsymbol{D})$ and $\nabla P(\boldsymbol{W}, \boldsymbol{D})$ are zero vector. This implies that the terms $\nabla M(\boldsymbol{W}, \boldsymbol{D})$ and $\nabla P(\boldsymbol{W}, \boldsymbol{D})$ are scalar multiples of each other, that is,

$$\nabla M(\boldsymbol{W}, \boldsymbol{D}) = -\lambda \nabla P(\boldsymbol{W}, \boldsymbol{D}). \tag{3.1}$$

Thus, condition (2) can be easily identified by means of inner product of the direction vectors of the two gradient vectors, $\nabla M(\boldsymbol{W}, \boldsymbol{D})$ and $\nabla P(\boldsymbol{W}, \boldsymbol{D})$. In this paper, the direction vector of a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ is defined by

$$\hat{\boldsymbol{x}} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \tag{3.2}$$

and the inner product between two direction vectors $\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)^{\mathrm{T}}$ and $\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^{\mathrm{T}}$ is defined by

$$\langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \rangle = \sum_{i=1}^{n} \hat{x}_i \hat{y}_i, \tag{3.3}$$

where the norm $\|\boldsymbol{x}\|$ is given by

$$\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{3.4}$$

Hence, the value of the inner product $\langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \rangle$ signifies the likelihood of getting stuck in a sub-optimal solution satisfying condition (2). In this project, the criterion of the stalling identification method is based on the value of the inner product $\langle \nabla \hat{M}, \nabla \hat{P} \rangle$. When the inner product is at its minimum value, -1, the training process stalls at a sub-optimal solution satisfying condition (2). Consequently, the mechanism of the stalling identification method is that the training process is classified as stalling when the inner product is less than a pre-selected threshold γ; otherwise, the training process is considered as not stalling.

3.2. λ selection schemes

Apart from the stalling identification method, the λ selection schemes are the other components which are of paramount importance in the ARPS method. The rationale behind the λ selection schemes is that when the training process is classified as not stalling, an appropriate λ is selected to guarantee the convergence of both $M(\boldsymbol{W}, \boldsymbol{D})$ and $P(\boldsymbol{W}, \boldsymbol{D})$ to

maximize the effect of the regularization method. While the network training is about to stall, another λ is chosen to assure the convergence of the term $M(\boldsymbol{W}, \boldsymbol{D})$ only. $P(\boldsymbol{W}, \boldsymbol{D})$ may not further converge, or even diverge slightly. In other words, the ARPS method, on the one hand, breaks the tendency of getting stuck in a sub-optimal solution satisfying condition (2) by means of changing λ. On the other hand, all the sub-optimal solutions of condition (2) disappear momentarily because the training process is, at that instance, switched into a non-regularized type training. Consequently, the network training may tunnel through the sub-optimal solution.

In order to make the above ideas work, a set of λ selection criteria has to be derived and are obtained by means of convergence analysis for the gradient descent type. The detailed derivation is summarized in Appendix A. We let the plausible range of λ for a particular regularized objective function be in the interval $(\lambda_{\min}, \lambda_{\max})$. The interval $(\lambda_{\min}, \lambda_{\max})$ is often determined by means of trial and error because the interval depends upon the nature of the training data and the regularized objective function. We derive a sufficient condition for the convergence of $M(\boldsymbol{W}, \boldsymbol{D})$. The change of $M(\boldsymbol{W}, \boldsymbol{D})$ is derived by

$$\Delta M = M(\boldsymbol{W} + \Delta \boldsymbol{W}, \boldsymbol{D}) - M(\boldsymbol{W}, \boldsymbol{D}). \tag{3.5}$$

Since the gradient descent training technique is used in this project, the update vector ΔW is proportional to ∇H, viz. $-\eta \nabla H$ where η is the learning rate. Using Taylor expansion, we have

$$\Delta M \approx \langle \nabla M, -\eta \nabla H \rangle. \tag{3.6}$$

Using the Lyapunov method, the sufficient condition for the convergence of the term $M(\boldsymbol{W}, \boldsymbol{D})$ is given by

$$\lambda \geqslant \frac{-\|\nabla M\|^2}{\langle \nabla M, \nabla P \rangle} \quad \text{and} \quad \lambda > 0. \tag{3.7}$$

Similarly, the sufficient condition for the convergence of the term $P(\boldsymbol{W}, \boldsymbol{D})$ is

$$\lambda \geqslant \frac{-\langle \nabla M, \nabla P \rangle}{\|\nabla P\|^2} \quad \text{and} \quad \lambda > 0. \tag{3.8}$$

In order to guarantee the convergence of the terms $M(\boldsymbol{W}, \boldsymbol{D})$ and $P(\boldsymbol{W}, \boldsymbol{D})$, Eqs. (3.9) and (3.10) are the criteria applicable to λ selection schemes A and B, respectively. In scheme A, the selection of λ is based on the following condition

$$0 < \max \left\{ \frac{-\langle \nabla M, \nabla P \rangle}{\|\nabla P\|^2}, \frac{-\|\nabla M\|^2}{\langle \nabla M, \nabla P \rangle} \right\} \leqslant \lambda, \tag{3.9}$$

when the inner product $\langle \nabla \hat{M}, \nabla \hat{P} \rangle$ is negative and greater than γ. When the inner product is greater than zero, in accordance with the conditions in Eqs. (3.7) and (3.8), a positive real number in $(\lambda_{\min}, \lambda_{\max})$ is theoretically a suitable choice of λ. In this project, the λ is set to be half of the $\lambda_{\max}$ when $\langle \nabla \hat{M}, \nabla \hat{P} \rangle$ is greater than zero. Besides, the λ selection scheme B is to assure the convergence of the $M(\boldsymbol{W}, \boldsymbol{D})$ only. The value of λ in scheme B can be selected from the following interval:

$$0 < \lambda \leqslant \min \left\{ \frac{-\langle \nabla M, \nabla P \rangle}{\|\nabla P\|^2}, \frac{-\|\nabla M\|^2}{\langle \nabla M, \nabla P \rangle} \right\}, \tag{3.10}$$

when the inner product is less than zero. The interval defined in Eq. (3.10) makes the value of λ sufficiently small so that the term $\nabla P(\boldsymbol{W}, \boldsymbol{D})$ will become negligible in the training process. The training process can be switched into a nonregularized type training based on the objective function $M(\boldsymbol{W}, \boldsymbol{D})$. In the λ selection scheme A, λ is computed by

$$\lambda_A = \min\left\{ \max\left\{ \frac{-\langle \nabla M, \nabla P\rangle}{\|\nabla P\|^2}, \frac{-\|\nabla M\|^2}{\langle \nabla M, \nabla P\rangle} \right\}, \lambda_{\max} \right\}, \tag{3.11}$$

when the inner product is less than zero. In the scheme B, λ is calculated by

$$\lambda_B = \frac{1}{2}\left(\max\{0, \lambda_{\min}\} + \min\left\{ \frac{-\langle \nabla M, \nabla P\rangle}{\|\nabla P\|^2}, \frac{-\|\nabla M\|^2}{\langle \nabla M, \nabla P\rangle} \right\} \right), \tag{3.12}$$

when the inner product is less than zero. Consequently, once the ARPS method is applied, the advantages of the regularization method are maximized and the problem of the sub-optimal solution satisfying condition (2) is eliminated within the plausible range of λ. Hereafter, the algorithm outline of our ARPS method is summarized as follows:

(1) To initialize $\boldsymbol{W}_0$ and λ_0;
(2) $\boldsymbol{W}_{k+1} = \boldsymbol{W}_k + \Delta \boldsymbol{W}_k$;
(3) If the training error is smaller than a presumed value, then stop;
(4) If $\langle \nabla \hat{M}, \nabla \hat{P}\rangle < \gamma$, then the training process is classified as "stalling" and jump to step (7);
(5) If $\langle \nabla \hat{M}, \nabla \hat{P}\rangle \geqslant 0$, then select λ_{k+1} to be $\lambda_{\max}/2$ and jump to step (2);
(6) Select λ_{k+1} based on Eq. (3.11) and jump to step (2);
(7) Select λ_{k+1} based on Eq. (3.12) and jump to step (2).

4. Simulation results

The proposed ARPS method was validated by applying it to two developed regularized type objective functions, namely weight decay ($P(\boldsymbol{W}, \boldsymbol{D}) = \sum_i w_i^2$) [2] and weight elimination ($P(\boldsymbol{W}, \boldsymbol{D}) = \sum_i (w_i^2/w_0^2)/(1 + w_i^2/w_0^2)$) [10]. Since the prediction of the sunspot series is regarded as a benchmark test, an example of noisy sunspot series sampled from the real world was used. In this study, the simulations were off-line and batch-mode based. The simulations were all run under a SUN Sparc 20 platform. In this paper, we illustrate that gradient descent type optimization over the regularized objective functions is able to tunnel through the sub-optimal solutions satisfying condition (2), when our proposed ARPS method is applied. To have a fair comparison, the same set of initial weight components were used throughout this study and each simulation ran the same number of iterations. The initial weight vector was randomized within the range between -1 and 1. Each simulation ran 30000 iterations and it is based on the condition with learning rate of 0.1 and momentum factor of 0.9.

The sunspot data (1700–1979) are divided into a training set (1700–1920) and two test sets, covering the periods of 1921–1955 (test 1) and 1956–1979 (test 2). In this example, the architecture of the BPnet is identical to that used by Weigend et al. [10], which has 12 inputs, 8 hidden units, and 1 output. The data of the sunspot series is normalized in the range between 0 and 1. The threshold γ is selected to be -0.6. The simulation

Table 1
Comparison of the results of the different methods

	RMS error					
	Weight decay			Weight elimination		
λ	Training set	Test set 1	Test set 2	Training set	Test set 1	Test set 2
0.001	0.34829	0.32012	0.28473	2.0747	2.0445	1.9710
0.01	0.17260	0.20029	0.30496	3.8184	3.7865	3.7106
0.1	0.18507	0.23013	0.36391	0.18456	0.22925	0.36241
1	0.23979	0.29365	0.44524	0.23936	0.29320	0.44474
ARPS	0.075137	0.072989	0.12655	0.064711	0.071126	0.13834

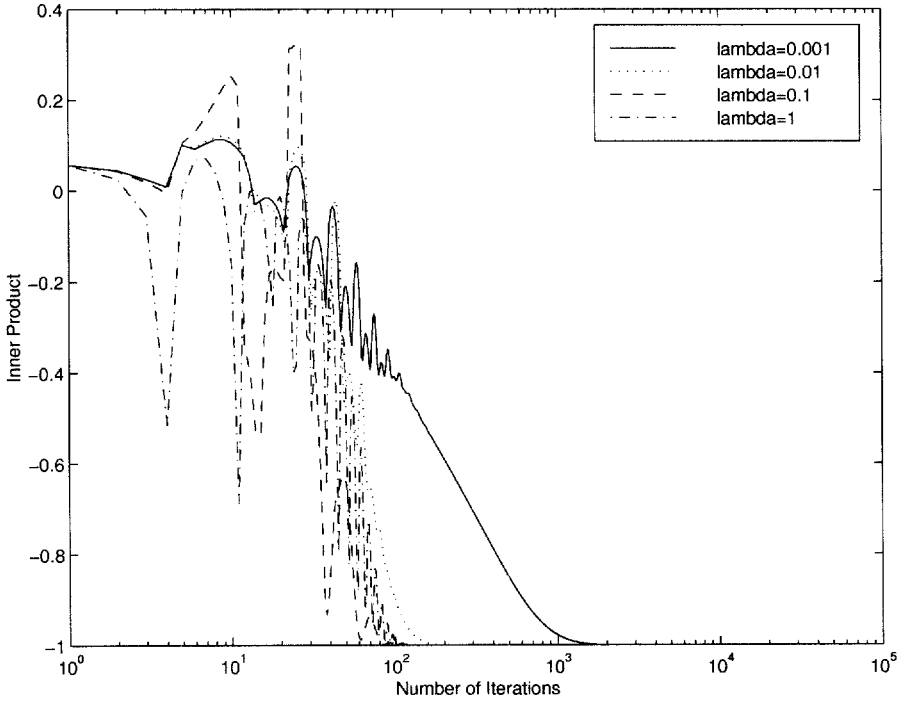

Fig. 1. The comparison of the change of the inner product $\langle \nabla \hat{M}, \nabla \hat{P} \rangle$ of weight decay method with different fixed λ.

results are summarized in Table 1 and Figs. 1–3. Figs. 1 and 2 illustrate that when different fixed λ were used, the value of the inner product $\langle \nabla \hat{M}, \nabla \hat{P} \rangle$ converged to -1 no matter what regularized objective function was used. In other words, if a fixed λ is used, the training process may easily stall at a sub-optimal solution of condition (2). Fig. 3 indicates that when the ARPS method was applied, the value of the inner product $\langle \nabla \hat{M}, \nabla \hat{P} \rangle$ did

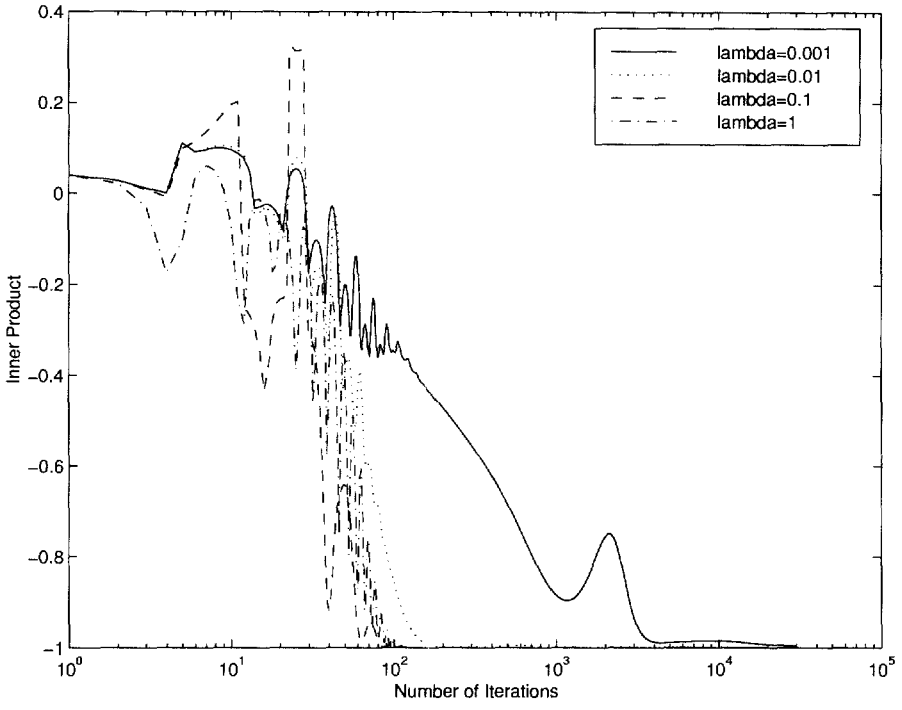

Fig. 2. The comparison of the change of the inner product $\langle \nabla \hat{M}, \nabla \hat{P} \rangle$ of weight elimination method with different fixed λ.

not converge to -1. The value of inner product appeared to approach the value of the threshold γ. These results substantiate that the proposed λ selection schemes A and B can effectively prevent the network training from getting stuck at a sub-optimal solution satisfying condition (2) and the stalling identification scheme can also detect the stalling in advance. Furthermore, Table 1 summarizes the root-mean-square errors of the simulations. The result enhancement due to the ARPS method is in terms of not only the training errors but also the two test errors. This corroborates that the ARPS method is capable of maximizing the effect of the regularization method in enhancing the generalization capability. The ARPS method enables the training process not to stall at the large number of sub-optimal solutions satisfying condition (2) although the network training might get stuck at a sub-optimal solution of condition (1).

5. Concluding remarks

A novel adaptive regularization parameter selection (ARPS) method was proposed to enhance the performance of the regularization method. The adaptive regularization parameter selection method enables a gradient descent type algorithm to tunnel through some of the undesired sub-optimal solutions on the composite error surface. This is achieved by changing the value of the regularization parameter before the training gets

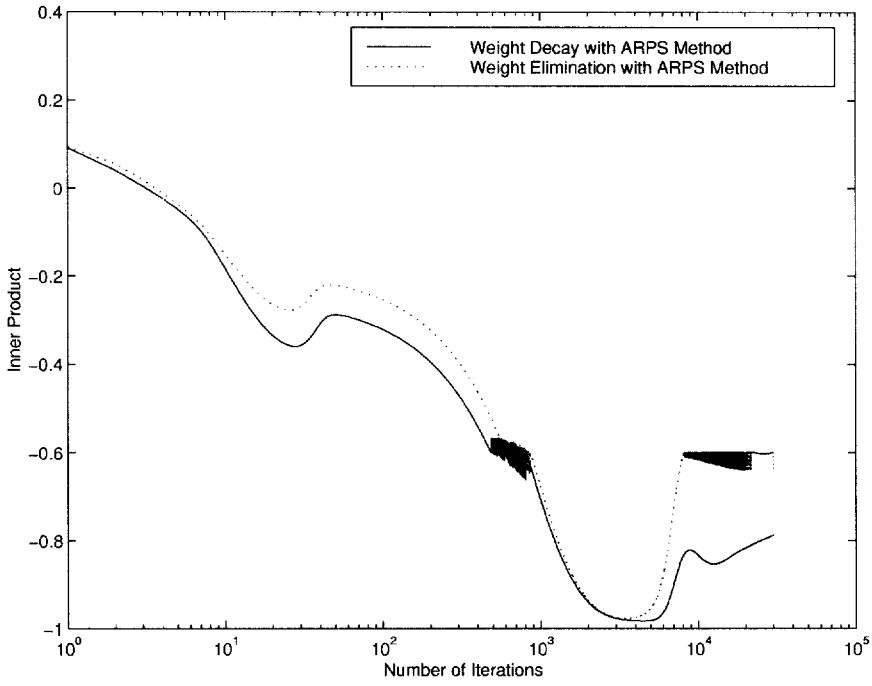

Fig. 3. The comparison of the change of the inner product $\langle \nabla \hat{M}, \nabla \hat{P} \rangle$ of weight decay and weight elimination methods with ARPS method.

stuck in a sub-optimal solution. The undesired solutions are introduced inherently from regularized objective functions. Hence, the proposed adaptive regularization parameter selection method is capable of enhancing the regularization method without getting stuck at the undesired sub-optimal solutions. The proposed method was validated by applying it to two regularization type objective functions. The obtained results substantiate that the proposed ARPS method can effectively prevent the network training from getting stuck at a sub-optimal solution satisfying condition (2).

Appendix A

A.1. The derivation of the sufficient conditions of the regularization parameter

We consider the regularized objective function is defined as

$$H(\boldsymbol{W}) = M(\boldsymbol{W}) + \lambda P(\boldsymbol{W}), \tag{A.1}$$

where λ is the regularization parameter and is positive. In this derivation, a gradient descent type training method is considered, viz.

$$\Delta \boldsymbol{W} = -\eta \nabla H, \tag{A.2}$$

where η is the learning factor.

Now, we consider the convergence of $M(\boldsymbol{W})$. After each iteration, the change of $M(\boldsymbol{W})$ is given by

$$\Delta M = M(\boldsymbol{W} + \Delta \boldsymbol{W}) - M(\boldsymbol{W}). \tag{A.3}$$

Since $\Delta \boldsymbol{W} = -\eta \nabla H$, using Taylor expansion, we have

$$\begin{aligned}\Delta M &\approx (M(\boldsymbol{W}) + \langle \nabla M(\boldsymbol{W}), \Delta \boldsymbol{W} \rangle) - M(\boldsymbol{W}) \\ &\approx \langle \nabla M(\boldsymbol{W}), -\eta \nabla H(\boldsymbol{W}) \rangle \\ &\approx -\eta \big(\langle \nabla M(\boldsymbol{W}), \nabla M(\boldsymbol{W}) \rangle + \lambda \langle \nabla M(\boldsymbol{W}), \nabla P(\boldsymbol{W}) \rangle \big).\end{aligned} \tag{A.4}$$

Using Lyapunov method, we have $\Delta M \leqslant 0$ and subsequently, we obtain the sufficient condition for the convergence of $M(\boldsymbol{W})$

$$\lambda \geqslant \frac{-\|\nabla M(\boldsymbol{W})\|^2}{\langle \nabla M(\boldsymbol{W}), \nabla P(\boldsymbol{W}) \rangle}. \tag{A.5}$$

The convergence of $P(\boldsymbol{W})$ is now considered. Similar to the case of $M(\boldsymbol{W})$, we have

$$\begin{aligned}\Delta P(\boldsymbol{W}) &= P(\boldsymbol{W} + \Delta \boldsymbol{W}) - P(\boldsymbol{W}) \\ &\approx \langle \nabla P(\boldsymbol{W}), \Delta \boldsymbol{W} \rangle \\ &\approx -\eta \big(\langle \nabla P(\boldsymbol{W}), \nabla M(\boldsymbol{W}) \rangle + \lambda \|\nabla P\|^2 \big).\end{aligned} \tag{A.6}$$

Consequently, the sufficient condition for the convergence of $P(\boldsymbol{W})$ is

$$\lambda \geqslant \frac{\langle \nabla P(\boldsymbol{W}), \nabla M(\boldsymbol{W}) \rangle}{-\|\nabla P(\boldsymbol{W})\|^2}. \tag{A.7}$$

References

[1] Y.S. Abu-Mostafa, Hints, Neural Computation 7 (4) (1995) 639–671.
[2] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.
[3] G. Cybenko, Approximation by superposition of a sigmoidal function, Math. Contr. Signal Syst. 2 (1989).
[4] G. Deco, W. Finnoff, H.G. Zimmermann, Unsupervised mutual information criterion for elimination of overtraining in supervised multilayer networks, Neural Computation 7 (1995) 86–107.
[5] K. Funahashi, On the approximate realization of continuous mappings by neural networks, Neural Networks 2 (1989) 183–192.
[6] S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma, Neural Computation 4 (1992) 1–58.
[7] S. Hochreiter, J. Schmidhuber, Flat minima, Neural Computation 9 (1997) 1–42.
[8] C.-T. Leung, T.W.S. Chow, Y.F. Yam, A least third-order cumulants objective function, Neural Process. Lett. 3 (2) (1996) 91–99.
[9] R. Setiono, A penalty-function approach for pruning feedforward network networks, Neural Computation 9 (1997) 185–204.
[10] A. Weigend, D. Rumelhart, B. Huberman, Generalization by weight elimination with application to forecasting, in: R.P. Lippman, J. Moody (Eds.), Advances in Neural Information Processing III, Morgan Kaufmann, San Mateo, CA, 1991, pp. 875–882.